Fictitious play applied to sequences of games and discounted stochastic games

Vrieze, O.J.; Tijs, S.H.

Published in:
International Journal of Game Theory

Publication date:
1982

1982

International Journal of Game Theory, Vol. 11, Issue 2, page 71–85.

Fictitious Play Applied to Sequences of Games and Discounted Stochastic Games

By *O.J. Vrieze* and *S.H. Tijs*, Nijmegen[1])

Abstract: In this paper, we show that the iterative method of Brown and Robinson, for solving a matrix game, is also applicable to a converging sequence of matrices, where the players choose at stage t a row and a column of the t-th matrix in the sequence. As an application of this result, we describe a new solution method for discounted stochastic games with finite state and action spaces.

1. Introduction

A long time ago, *Brown* [1949, 1951] suggested a method, called fictitious play, for solving a matrix game and *Robinson* [1950] proved the validity of that method. Rightly, this beautiful result was included in many books on game theory, such as *McKinsey* [1952], *Luce/Raiffa* [1957], *Karlin* [1959], *Dresher* [1961], *Owen* [1968] and *Rauhut/Schmitz/Zachow* [1979]. *Shapiro* [1958] provided for the Brown-Robinson scheme an a priori estimate of the rate of convergence. Extensions of the Brown-Robinson method to infinite zero-sum games were given in *Danskin* [1954] and *Van den Akker* [1976]; and an extension to 2×2-bimatrix games in *Miyasawa* [1961]. In *Shapley* [1964], it was shown by examples, that the natural generalization of the Robinson theorem to arbitrary bimatrix games is not valid. A systematic study of this phenomenon was done by *Rosenmüller* [1971].

The method of fictitious play of Brown and Robinson can be seen as an infinite stage learning process, corresponding to a matrix game A, where at each stage the players choose a pure action which, among the pure actions, would have been the best against the total of past choices of the other player.

The purpose of this paper is to extend the ideas of Brown and Robinson into two directions. Firstly, we extend them in section 2 to a situation, in which the matrix game A, which is played is not exactly known in advance, but where at each stage $t \in \mathbf{N}$ an approximation $A(t)$ is given, and where $\lim_{t \to \infty} A(t) = A$. Secondly, in section 3 we describe an iterative method for solving a discounted stochastic game, with finite state and action spaces. The paper concludes in section 4 with some remarks.

---

[1]) Ir. *O.J. Vrieze*, and Dr. *S.H. Tijs*, Department of Mathematics, Catholic University, Toernooiveld, NL–6525 ED Nijmegen.
0020-7276/82/020071-85 $2.50

2. Fictitious Play in a Not Perfectly Known Situation

In this section we show that the iterative method of solving a matrix game A of *Brown* [1949] and *Robinson* [1950] can be modified to a situation, where at each step t of the iteration procedure, not A but an approximation $A(t)$ is used, known at step t. Hence, we consider a converging sequence

$$A(1), A(2), A(3), \ldots$$

of $m \times n$-matrices with $\lim_{t\to\infty} A(t) = A$ i.e.

$$\lim_{t\to\infty} a_{ij}(t) = a_{ij} \text{ for all } i \in \{1,2,\ldots,m\} \text{ and } j \in \{1,2,\ldots,n\}.$$

In the following, for each $\epsilon > 0$, we denote by $t(\epsilon)$ the smallest integer, such that

$$|a_{ij}(t) - a_{ij}| \leqslant \epsilon \text{ for all } i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\} \text{ and } t \geqslant t(\epsilon). \qquad (2.1)$$

We also use the following notation.

For a vector $X = (x_1, x_2, \ldots, x_r) \in \mathbf{R}^r$, $\max\{x_1,\ldots,x_r\}$ is denoted by $\max X$ and $\min\{x_1,\ldots,x_r\}$ by $\min X$. For $A(t)$ and A the j-th column is denoted by $C_j(t)$ and C_j, respectively; and the i-th row by $R_i(t)$ and R_i. Moreover, $\mathbf{N}_0 = \mathbf{N} \cup \{0\}$ and

$$M = \sup\{|a_{ij}(t)| : i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}, t \in \mathbf{N}\} \in \mathbf{R}.$$

Definition 2.1: We call a pair of sequences

$$X(0), X(1), X(2), \ldots \quad \text{in } \mathbf{R}^n,$$
$$Y(0), Y(1), Y(2), \ldots \quad \text{in } \mathbf{R}^m$$

a *vector system* for the sequence $\langle A(t): t \in \mathbf{N}\rangle$, if

(V.1) $\min X(0) = \max Y(0)$,

and if for each $t \in \mathbf{N}$

(V.2) $X(t) = X(t-1) + R_{i(t)}(t)$, where $i(t) \in \{1,\ldots,m\}$ satisfies

$$y_{i(t)}(t-1) = \max Y(t-1),$$

(V.3) $Y(t) = Y(t-1) + C_{j(t)}(t)$, where $j(t) \in \{1,\ldots,n\}$ satisfies

$$x_{j(t)}(t-1) = \min X(t-1)$$

[$y_r(t)$ is the r-th coordinate of $Y(t)$].

It is obvious, how such a vector system can be formed, recursively, from given $X(0)$ and $Y(0)$, satisfying (V.1). We want to show that

$$\lim_{t\to\infty} t^{-1} \max Y(t) = \lim_{t\to\infty} t^{-1} \min X(t) = \text{val}(A),$$

where val (A) is the value of the matrix game A. In proving this, we follow as far as possible a similar line as in the paper of *Robinson* [1950], where the situation is studied, in which $A(t) = A$ for each $t \in \mathbf{N}$.

Lemma 2.1: $\limsup_{t\to\infty} t^{-1} \min X(t) \leqslant \text{val}(A) \leqslant \liminf_{t\to\infty} t^{-1} \max Y(t)$.

Proof: Take $\epsilon > 0$ and $t > t(\epsilon)$. Then

$$X(t) = X(t(\epsilon)) + \sum_{\tau=t(\epsilon)+1}^{t} R_{i(\tau)}(\tau) \leqslant X(t(\epsilon)) + \sum_{\tau=t(\epsilon)+1}^{t} R_{i(\tau)} + \epsilon(t - t(\epsilon))\, 1_n,$$

where 1_n is the vector in $\mathbf{R}^n$, with all coordinates equal to 1. Let $a_i(t)$ be the number of times that $i \in \{1, \ldots, m\}$ appears in the sequence $\langle i(t(\epsilon)+1), i(t(\epsilon)+2), \ldots, i(t)\rangle$.

Then $\pi(t) = (t - t(\epsilon))^{-1} (a_1(t), \ldots, a_m(t))$ is a mixed action for player 1 in the matrix game A with $\pi(t) A = (t - t(\epsilon))^{-1} \sum_{\tau=t(\epsilon)+1}^{t} R_{i(\tau)}$. Hence,

$$X(t) \leqslant X(t(\epsilon)) + (t - t(\epsilon))\, \pi(t) A + \epsilon(t - t(\epsilon))\, 1_n,$$

which implies that

$$t^{-1} \min X(t) \leqslant t^{-1} \max X(t(\epsilon)) + t^{-1}(t - t(\epsilon)) \min \pi(t) A + \epsilon t^{-1}(t - t(\epsilon))$$

$$\leqslant t^{-1} \max X(t(\epsilon)) + t^{-1}(t - t(\epsilon)) \text{val}(A) + \epsilon t^{-1}(t - t(\epsilon)).$$

Hence, $\limsup_{t\to\infty} t^{-1} \min X(t) \leqslant \text{val}(A) + \epsilon$ for each $\epsilon > 0$, from which the first inequality in the lemma follows. The second inequality can be proved similarly. □

Definition 2.2: If $\langle X(t), Y(t): t \in \mathbf{N}_0 \rangle$ is a vector system for the sequence of matrices $\langle A(t): t \in \mathbf{N}\rangle$, then we say that the i-th pure action of player 1 is *eligible* for the vector system in the interval $[t, t']$, if there exists a $t_1 \in [t, t']$, such that $y_i(t_1) = \max Y(t_1)$. Eligibility of a pure action j for player 2, is defined analogously.

Lemma 2.2: If, for $s, t \in \mathbf{N}$, *all pure actions of both players are eligible in* $[s, s+t]$, *then*

$$\max X(s+t) - \min X(s+t) \leqslant 2tM,$$

$$\max Y(s+t) - \min Y(s+t) \leqslant 2tM.$$

Proof: The lemma follows by modifying in an obvious way the proof of lemma 2 in *Robinson* [1950], using the definition of M. □

Lemma 2.3: If, for an $s \geqslant t\ (\epsilon)$, *all pure actions of both players are eligible in* $[s, s+t]$, *then*

$$\max Y(s+t) - \min X(s+t) \leqslant 4tM + 2\epsilon(s+t) + 2Mt\ (\epsilon). \tag{2.2}$$

Proof: In view of lemma 2.2:

$$\max Y(s+t) - \min X(s+t) \leqslant 4tM + \min Y(s+t) - \max X(s+t). \tag{2.3}$$

Let $\pi(s+t)$ be the mixed action for player 1, as defined in the proof of lemma 2.1. The inequality

$$X(s+t) \geqslant X(t(\epsilon)) + (s+t-t(\epsilon))\,\pi(s+t)\,A - \epsilon(s+t-t(\epsilon))\,1_n$$

implies then:

$$\max X(s+t) \geqslant \min X(t(\epsilon)) + (s+t-t(\epsilon))\,\mathrm{val}(A^t) - \epsilon(s+t-t(\epsilon)) \tag{2.4}$$

where A^t is the transpose of the matrix A. Similarly, one can show that

$$\min Y(s+t) \leqslant \max\ Y(t(\epsilon)) + (s+t-t(\epsilon))\,\mathrm{val}(A^t) + \epsilon(s+t-t(\epsilon)). \tag{2.5}$$

Note further, that by (V.1) and the definition of M:

$$\max Y(t(\epsilon)) - \min X(t(\epsilon)) \leqslant 2Mt\ (\epsilon). \tag{2.6}$$

Combining the inequalities (2.3) – (2.6), yields (2.2). □

For a matrix B, we denote by B^{-i} (B_{-j}) the matrix, which is obtained from B by deleting the i-th row (the j-th column); for a vector $Y = (y_1, y_2, \ldots, y_m) \in \mathbf{R}^m$, let Y^{-i} be the vector $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$.

Lemma 2.4: Let $\langle X(t), Y(t): t \in \mathbf{N}_0 \rangle$ *be a vector system for the sequence* $\langle A(t): t \in \mathbf{N} \rangle$ *of* $m \times n$*-matrices, converging to* A. *Suppose that, in the interval* $[s, s+t_0]$, *pure action* i *of player* 1 *is not eligible for the vector system. For* $\tau \in \{0, 1, \ldots, t_0\}$, *let*

$$X'(\tau) = X(s+\tau) + (\max Y(s) - \min X(s))\, 1_n,\ Y'(\tau) = Y^{-i}(s+\tau).$$

Then $\langle X'(\tau), Y'(\tau): \tau \in \{0,1,\ldots,t_0\}\rangle$ *is the start of a vector system for the converging sequence of* $(m-1) \times n$*-matrices* $\langle A^{-i}(s+t): t \in \mathbf{N}\rangle$. *Furthermore,*

$$\max Y(s+t_0) - \min X(s+t_0) = (\max Y'(t_0) - \min X'(t_0)) + (\max Y(s) - \min Y(s)).$$

Proof: Obviously, $\min X'(0) = \max Y'(0)$, since pure action i is ineligible in s. Because $\langle X(t), Y(t)\rangle$ is a vector system and i is not eligible in $[s, s+t_0]$, for $i'(\tau) = i(s+\tau) \in \{i, \ldots, i-1, i+1, \ldots, m\}$ and $j'(\tau) = j(s+\tau) \in \{1, \ldots, n\}$, $(\tau = 1, \ldots, t_0)$, we have

$$X(s+\tau) = X(s+\tau-1) + R_{i'(\tau)}(s+\tau),\ y_{i'(\tau)}(s+\tau-1) = \max Y(s+\tau-1)$$

$$Y(s+\tau) = Y(s+\tau-1) + C_{j'(\tau)}(s+\tau),\ x_{j'(\tau)}(s+\tau-1) = \min X(s+\tau-1).$$

These equalities imply for $\tau \in [1, t_0]$:

$$X'(\tau) = X'(\tau-1) + R_{i'(\tau)}(s+\tau), y'_{i'(\tau)}(\tau-1) = \max Y'(\tau-1),$$

$$Y'(\tau) = Y'(\tau-1) + C^{-i}_{j'(\tau)}(s+\tau),\ x'_{j'(\tau)}(\tau-1) = \min X'(\tau-1).$$

Consequently, the first assertion in the lemma is proved. Furthermore,

$$\max Y(s+t_0) - \min X(s+t_0) = \max Y'(t_0) - (\min X'(t_0) - (\max Y(s) - \min Y(s))),$$

which finishes the proof. □

In an obvious way, one can formulate a "player 2 version" of lemma 2.4, where a non-eligible action j of player 2 plays a role. Both versions will be used in the proof of the following lemma.

Lemma 2.5: Let $\langle A(t): t \in \mathbf{N}\rangle$ *be a converging sequence of matrices and let* $\epsilon > 0$. *Then, there exists a non-negative real number* $T(\epsilon)$, *such that for each* $\tau \in \mathbf{N}_0$ *and each vector system* $\langle X(t), Y(t): t \in \mathbf{N}_0\rangle$ *of the sequence* $\langle A(t+\tau): t \in \mathbf{N}\rangle$, *we have*

$$\max Y(t) - \min X(t) \leqslant \epsilon t, \text{ for all } t \geqslant T(\epsilon).$$

Proof: The lemma will be proved by induction to the size of the matrices (number of rows + number of columns) in the sequence. For sequences of 1×1-matrices (of size 2) we can take $T(\epsilon) = 0$ since

$$Y(t) = X(t) \in \mathbf{R} \quad \text{for all } t \in \mathbf{N}_0.$$

Now suppose that the statement is true for converging sequences of $k \times \ell$-matrices, with size $k + \ell < m + n$.

Let $\langle A(t): t \in \mathbf{N}\rangle$ be a converging sequence of $m \times n$-matrices and let $\epsilon > 0$. By

applying the induction hypotheses to the finite number of converging sequences $\langle A^{-i}(t): t \in \mathbf{N}\rangle$ $(i \in \{1, \ldots, m\})$ and $\langle A_{-j}(t): t \in \mathbf{N}\rangle$ $(j \in \{1, \ldots, n\})$ of size $m + n - 1$, we may conclude that there exists a $\hat{T}(\epsilon)$ such that for each $\tau \in \mathbf{N}_0$ and each vector system $\langle X'(t), Y'(t)\rangle$ of one of the sequences $\langle A^{-i}(t+\tau): t \in \mathbf{N}\rangle$ or $\langle A_{-j}(t+\tau): t \in \mathbf{N}\rangle$ we have

$$\max Y'(t) - \min X'(t) \leqslant \epsilon t \text{ for all } t \geqslant \hat{T}(\epsilon). \tag{2.7}$$

Take a $\tau \in \mathbf{N}_0$ and an arbitrary vector system $\langle X(t), Y(t): t \in \mathbf{N}_0\rangle$ of $\langle A(t+\tau): t \in \mathbf{N}\rangle$. Let $t \geqslant t(\epsilon) + \hat{T}(\epsilon)$. Then there is a $q \in \mathbf{N}$ and $r \in [0,1)$ such that $t = t(\epsilon) + (r+q)\,T(\epsilon)$. We distinguish two cases:

Case 1: Suppose that there is an integer $p \in \{1,2, \ldots, q\}$, such that all actions of both players are eligible in the interval

$$[t(\epsilon) + (r + (p-1))\,\hat{T}(\epsilon),\ t(\epsilon) + (r+p)\,\hat{T}(\epsilon)].$$

Let z be the largest integer in $\{1,2, \ldots, q\}$ with that property. Then lemma 2.4 (and the player 2 version of lemma 2.4) may be repeatedly applied with $[t(\epsilon) + + (r+k-1)\,\hat{T}(\epsilon),\ t(\epsilon) + (r+k)\,\hat{T}(\epsilon)]$ in the role of $[s, s+t_0]$ for $k = = z+1, \ldots, q$, using the fact that in each of these intervals at least one action for one of the players is not eligible, yielding in view of (2.7):

$$\max Y(t) - \min X(t) = (\max Y(t(\epsilon) + (r+z)\,\hat{T}(\epsilon)) - \min X(t(\epsilon) + (r+z)\,\hat{T}(\epsilon)) + (q-z) \in \hat{T}(\epsilon). \tag{2.8}$$

Since all actions are eligible in the interval $[t(\epsilon) + (r+z-1)\,\hat{T}(\epsilon),\ t(\epsilon) + + (r+z)\,\hat{T}(\epsilon)]$, the first term in the right hand side of (2.8) is by lemma 2.3, at most equal to

$$4\hat{T}(\epsilon)\,M + 2\epsilon\,(t(\epsilon) + (r+z)\,\hat{T}(\epsilon)) + 2Mt(\epsilon) \leqslant 2\epsilon t + 4\hat{T}(\epsilon)\,M + 2Mt(\epsilon)$$

and the second term is at most ϵt. Hence

$$\max Y(t) - \min X(t) \leqslant 3\epsilon t + \alpha$$

where $\alpha = M(4\hat{T}(\epsilon) + 2t(\epsilon))$. Consequently,

$$\max Y(t) - \min X(t) \leqslant 4\epsilon t, \text{ if } t \geqslant \max\{t(\epsilon) + \hat{T}(\epsilon), \epsilon^{-1}\alpha\}. \tag{2.9}$$

Case 2: If there is no such an integer p with the property, described in case 1, then lemma 2.4 (and its player 2 version) can be applied q times, yielding, in view of (2.7):

$$\max Y(t) - \min X(t) \leqslant \max Y(t(\epsilon) + r\hat{T}(\epsilon)) - \min X(t(\epsilon) + r\hat{T}(\epsilon)) + + q\,\epsilon\,\hat{T}(\epsilon) \leqslant 2M(t(\epsilon) + \hat{T}(\epsilon)) + \epsilon t \leqslant \epsilon t + \alpha$$

which implies that

$$\max Y(t) - \min X(t) \leqslant 4\epsilon t \text{ if } t \geqslant \max\left\{t(\epsilon) + \hat{T}(\epsilon), \frac{1}{3}\epsilon^{-1}\alpha\right\}. \tag{2.10}$$

Combining the two cases, we have, by (2.9) and (2.10):

$$\max Y(t) - \min X(t) \leqslant 4\epsilon t \text{ for all } t \geqslant T(4\epsilon),$$

if we take $T(4\epsilon) = \max\{t(\epsilon) + \hat{T}(\epsilon), \epsilon^{-1}\alpha\}$. This completes the proof of the induction step, since $T(4\epsilon)$ does not depend on τ and on the vector system $\langle X(t), Y(t): t \in \mathbf{N}_0\rangle$. □

Theorem 2.1: Let $\langle A(t): t \in \mathbf{N}\rangle$ *be a sequence of matrices with* $\lim_{t\to\infty} A(t) = A$. Then, for each vector system $\langle X(t), Y(t): t \in \mathbf{N}_0 >$ *of this sequence, we have*

$$\lim_{t\to\infty} t^{-1} \max Y(t) = \lim_{t\to\infty} t^{-1} \min X(t) = \text{val}(A).$$

Proof: The theorem is a direct consequence of the lemmas 2.1 and 2.5. □

For a vector system $\langle X(t), Y(t): t \in \mathbf{N}_0\rangle$ of the sequence $\langle A(t): t \in \mathbf{N}\rangle$, converging to A, let for each $t \in \mathbf{N}_0$, $\hat{\pi}(t)$ and $\hat{\rho}(t)$ be the mixed actions of players 1 and 2, respectively, where $t\hat{\pi}_i(t)$ equals the number of times that i appears in $\langle i(1), i(2), \ldots, i(t)\rangle$ and $t\hat{\rho}_j(t)$ the number of times that j appears in $\langle j(1), j(2), \ldots, j(t)\rangle$.

Theorem 2.2: Each limit point of the sequence $\langle\hat{\pi}(t): t \in \mathbf{N}\rangle$ *is an optimal mixed action of player 1 in the matrix game A. Each limit point of* $\langle\hat{\rho}(t): t \in \mathbf{N}\rangle$ *is optimal for player 2 in A.*

Proof: We only prove the first assertion. Let π^* be a limit point of $\langle\hat{\pi}(t): t \in \mathbf{N}\rangle$. Without loss of generality, we suppose that $\lim_{t\to\infty} \hat{\pi}(t) = \pi^*$. (Otherwise, look at a subsequence.) Let $\epsilon > 0$ and let $t > t(\epsilon)$. Then

$$X(t) = X(0) + \sum_{\tau=1}^{t(\epsilon)} R_{i(\tau)}(\tau) + \sum_{\tau=t(\epsilon)+1}^{t} R_{i(\tau)}(\tau)$$

$$\leqslant X(0) + \sum_{\tau=1}^{t} R_{i(\tau)} + 2Mt(\epsilon)\,1_n + \epsilon(t - t(\epsilon))\,1_n$$

$$= X(0) + t\hat{\pi}(t)\,A + 2Mt(\epsilon)\,1_n + \epsilon(t - t(\epsilon))1_n.$$

From this inequality follows, by taking limits:

$$\lim_{t\to\infty} \min t^{-1} X(t) \leqslant \min \pi^* A + \epsilon, \text{ for each } \epsilon > 0.$$

But then by theorem 2.1,

$$\text{val}(A) \leqslant \min \pi^* A \leqslant \max_{\pi} \min \pi A = \text{val}(A).$$

Consequently, π^* is optimal for player 1 in the matrix game A. □

Remark 2.1: It is well-known [cf. *Karlin*, 74–76], that the set $U(m, n)$ of $m \times n$-matrix games with unique optimal actions for both players, is an open and dense subset of the set of all $m \times n$-matrix games. If we have a sequence $\langle A(t): t \in \mathbf{N}\rangle$, converging to an $A \in U_{m\times n}$ then the sequence $\langle \hat{\pi}(t): t \in \mathbf{N}\rangle$, appearing in theorem 2.2 converges, and the limit equals the unique optimal action of player 1 in the game A.

3. An Algorithm for Solving Discounted Stochastic Games

In this section, we present a new method for solving discounted stochastic games with finite state and action spaces. The algorithms, proposed till now, are all based on the contraction property of the value map, going back to the paper of *Shapley* [1953]. For a survey in this area, see *Van der Wal* [1977]. These algorithms have in common, that although the rate of convergence is fast (geometric), in each iteration step much work must be done: N linear programming problems have to be solved, where N is the number of states of the stochastic game. In the method, which we propose here, the convergence rate is relatively slow, but the amount of labor in each iteration step is also very small, namely N times searching the maximum coordinate of a vector and N times searching the minimum coordinate of a vector.

Our algorithm is based on the results in section 2, and can therefore be seen as an extension of the Brown-Robinson procedure to discounted stochastic games.

Before describing our algorithm, we recall some well-known facts, concerning stochastic games.

A *discounted* (two-person zero-sum) *stochastic game* can be characterized by a five tuple

$$\Gamma = \langle S, \{A_{mk}: k \in S, m \in \{1,2\}\}, r, P, \beta\rangle$$

where

(i) $S = \{1,2,\ldots,N\}$, called the *state space*,

(ii) $A_{mk} = \{1,2,\ldots,n_m(k)\}$, called the *pure action space for* player m in state k, $k \in \{1,2\}$,

(iii) r is a real-valued function, defined on the set of triples $T = \{(k, i, j): k \in S, i \in A_{1k}, j \in A_{2k}\}$, called the *reward function*,

(iv) P is a map from T into the set $\mathcal{P}(S)$ of probability measures on S, called the *transition probability map*,

(v) β is a positive real number smaller than 1, called the *discount factor*.

Such a stochastic game corresponds to a dynamic system with state space S, which is governed by two players 1, 2 with opposite interests, who at the decision moments (*stages*) 0,1,2, . . . observe the state of the system and then influence the development of the system by choosing one of their actions available at the observed state (mixed actions are allowed). If at stage t the system is in state k, and if player 1 selects action $i \in A_{ik}$ and player 2 action $j \in A_{2k}$, then two things happen.

1) player 1 obtains an immediate reward $r\,(k, i, j)$ from player 2,
2) the system moves with probability $P\,(k, i, j)\,\{\ell\}$ – which we shall denote in the following by $p\,(\ell \mid k, i, j)$ – to state $\ell \in S$, which is observed at stage $t + 1$.

It is assumed that a reward r for a player at stage t has worth $\beta^t r$ at stage 0; $\beta^t r$ is called the *discounted reward*. We suppose that player 1 (player 2) wants to maximize (minimize) the total discounted expected reward.

A *mixed action* for player 1 in state k is a probability vector $\pi\,(k) = (\pi_1\,(k), \pi_2\,(k), \ldots, \pi_{n_1(k)}\,(k))$ i.e. $\pi_i\,(k) \geqslant 0$ for all $i \in A_{1k}$ and $\sum_{i \in A_{1k}} \pi_i\,(k) = 1$. A *stationary strategy* for player 1 in the game Γ is an N-tuple $\pi = (\pi\,(1), \pi\,(2), \ldots, \pi\,(N))$, where the k-th coordinate $\pi\,(k)$ is a mixed action for player 1 in state k. Using a stationary strategy π means, that in all stages, where the system is in state k, player 1 uses his mixed action $\pi\,(k)$. For a stationary strategy π for player 1 and a stationary strategy ρ for player 2, we look at

$$r\,(\pi, \rho) = (r_1\,(\pi\,(1), \rho\,(1)), r_2\,(\pi\,(2), \rho\,(2)), \ldots, r_N\,(\pi\,(N), \rho\,(N))) \in \mathbf{R}^N,$$

where

$$r_k\,(\pi\,(k), \rho\,(k)) = \sum_{i=1}^{n_1(k)} \sum_{j=1}^{n_2(k)} \pi_i\,(k)\,\rho_j\,(k)\,r\,(k, i, j)$$

is the expected immediate reward for player 1 in state k, if the mixed actions $\pi\,(k)$ and $\rho\,(k)$ are used. Furthermore, let $P\,(\pi, \rho)$ be the $N \times N$-matrix with in the (k, ℓ)-th cell the real number

$$p\,(\ell \mid k, \pi\,(k), \rho\,(k)) = \sum_{i=1}^{n_1(k)} \sum_{j=1}^{n_2(k)} \pi_i\,(k)\,\rho_j\,(k)\,p\,(\ell \mid k, i, j),$$

which is the expected probability that the system at the next stage is in state ℓ, given that the system is presently in state k, and given that the players use $\pi\,(k)$ and $\rho\,(k)$, respectively, in state k. The total β-discounted expected payoff for player 1, if the system starts in state k and if the stationary strategies π and ρ are used, is now given by the k-th coordinate of the vector $\sum_{t=0}^{\infty} \beta^t\,P\,(\pi, \rho)^t r\,(\pi, \rho)$ in $\mathbf{R}^N$. This vector will be denoted by $v\,(\pi, \rho) = (v_1\,(\pi, \rho), v_2\,(\pi, \rho), \ldots, v_N\,(\pi, \rho))$, in the sequel. In *Shapley* [1953], where stochastic games were introduced, the following lemma was proved.

Lemma 3.1: *Each discounted two-person zero-sum stochastic game* Γ *(with finite state and action spaces) has a value*

$$V(\Gamma) = \sup_{\pi} \inf_{\rho} v(\pi, \rho) = \inf_{\rho} \sup_{\pi} v(\pi, \rho) \in \mathbf{R}^N$$

and for the players 1 and 2 there exist optimal stationary strategies π^* *and* ρ^* (i.e. $V(\Gamma) = \inf v(\pi^*, \rho) = \sup v(\pi, \rho^*)$).

Furthermore, a stationary strategy $\pi(\rho)$ *is optimal for player 1 (player 2) iff for each* $k \in S$, *the mixed action* $\pi(k)$ $(\rho(k))$ *is an optimal action for player 1 (player 2) in the* $n_1(k) \times n_2(k)$*-matrix game*

$$[r(k, i, j) + \beta \sum_{\ell=1}^{N} p(\ell \mid k, i, j) V_\ell(\Gamma)]_{i=1, j=1}^{n_1(k), n_2(k)},$$

and, moreover, this matrix game has value $V_k(\Gamma)$.

Let us introduce here a useful notation. For a $W \in \mathbf{R}^N$ and a mixed action $\pi(k)$ for player 1 in state k, we denote by $(r + \beta PW)_{\pi(k)}$, the vector in $\mathbf{R}^{n_2(k)}$ with j-th coordinate

$$r(k, \pi(k), j) + \beta \sum_{\ell=1}^{N} p(\ell \mid k, \pi(k), j) W_\ell.$$

Similarly, for a mixed action $\rho(k)$ of player 2, $(r + \beta PW)_{\rho(k)} \in \mathbf{R}^{n_1(k)}$ is the vector with

$$r(k, i, \rho(k)) + \beta \sum_{\ell=1}^{N} p(\ell \mid k, i, \rho(k)) W_\ell$$

as its i-th coordinate.

We now describe our *algorithm*. We define recursively a sequence

$$W(0), W(1), W(2), \ldots \in \mathbf{R}^N$$

and, for each $k \in S$, sequences

$$X(k, 0), X(k, 1), X(k, 2), \ldots \in \mathbf{R}^{n_2(k)}$$

$$Y(k, 0), Y(k, 1), Y(k, 2), \ldots \in \mathbf{R}^{n_1(k)}$$

as follows. For each $k \in S$, choose $X(k, 0) \in \mathbf{R}^{n_2(k)}$, $Y(k, 0) \in \mathbf{R}^{n_1(k)}$, such that

$$\min X(k, 0) = \max Y(k, 0) \text{ and } \min Y(k, 0) \geqslant V_k(\Gamma),$$

and let, for each $k \in S$, the k-th coordinate $W_k(0)$ of $W(0)$ be equal to $\max Y(k, 0)$. [Note that $V_k(\Gamma) \leqslant (1 - \beta)^{-1} \max \{r(k, i, j): (k, i, j) \in T\}$.] Assume that for $t \in \mathbf{N}$,

$k \in S$ the vectors $W(t-1)$, $X(k, t-1)$ and $Y(k, t-1)$ are defined. Take $i(k, t) \in A_{1k}$ and $j(k, t) \in A_{2k}$, such that $y_{i(k,t)}(k, t-1) = \max Y(k, t-1)$ and $x_{j(k,t)}(k, t-1) = \min X(k, t-1)$. Now set for each $k \in S$:

(S.1) $W_k(t) = \min \{\max t^{-1} Y(k, t-1), W_k(t-1)\}$,

(S.2) $X(k, t) = X(k, t-1) + (r + \beta PW(t))_{i(k,t)}$,

(S.3) $Y(k, t) = Y(k, t-1) + (r + \beta PW(t))_{j(k,t)}$.

For each $t \in \mathbf{N}$, let $\hat{\pi}(k, t)$ be the mixed action for player 1, in state k with i-th coordinate $\hat{\pi}_i(k, t)$ equal to $t^{-1}c_i$, where c_i is the number of elements in the sequence $\langle i(k, 1), \ldots, i(k, t)\rangle$, equal to i. Similarly, $\hat{\rho}(k, t)$ is the mixed action for player 2 in state k, with $t\hat{\rho}_j(k, t)$ equal to the number of times that j appears in the sequence $\langle j(k, 1), \ldots, j(k, t)\rangle$. Then the following two theorems hold.

Theorem 3.1: For each $k \in S$:

$$\lim_{t\to\infty} t^{-1} \max Y(k, t) = \lim_{t\to\infty} t^{-1} \min X(k, t) = \lim_{t\to\infty} W_k(t) = V_k(\Gamma).$$

Theorem 3.2: For each $k \in S$, let $\hat{\pi}(k)$ be a limit point of the sequence $\langle \hat{\pi}(k, t): t \in \mathbf{N}\rangle$ and $\hat{\rho}(k)$ a limit point of $\langle \hat{\rho}(k, t): t \in \mathbf{N}\rangle$. Then $\hat{\pi} = (\hat{\pi}(1), \hat{\pi}(2), \ldots, \hat{\pi}(N))$ and $\hat{\rho} = (\hat{\rho}(1), \ldots, \hat{\rho}(N))$ are optimal stationary strategies for player 1 and player 2, respectively, in the stochastic game Γ.

To prove these theorems, we need the following lemma.

Lemma 3.2: Notation as above. $\operatorname{Lim}_{t\to\infty} W(t)$ *exists.*

Proof: By definition of $W_k(t)$, for each $k \in S$, the sequence $W_k(0), W_k(1), W_k(2), \ldots$ is a decreasing sequence. Hence, we only have to show that the sequence is lower bounded. We prove by induction that for each $t \in \mathbf{N}_0$:

$$W_k(t) \geqslant V_k(\Gamma) \text{ for each } k \in S. \tag{3.1}$$

For $t = 0$, we have

$$W_k(0) = \max Y(k, 0) \geqslant \min Y(k, 0) \geqslant V_k(\Gamma) \text{ for each } k \in S.$$

Suppose, that $W_k(\tau) \geqslant V_k(\Gamma)$ for all $\tau \leqslant t \in \mathbf{N}_0$ and $k \in S$. Then

$$Y(k, t) = Y(k, 0) + \sum_{\tau=1}^{t} (r + \beta PW(\tau))_{j(k,\tau)}$$

$$\geqslant V_k(\Gamma)\, 1_{n_1(k)} + \sum_{\tau=1}^{t} (r + \beta PV(\Gamma))_{j(k,\tau)}.$$

Now, for the mixed action $\hat{\rho}(k, t)$, introduced before, we have

$$\sum_{\tau=1}^{t} (r + \beta PV(\Gamma))_{j(k,\tau)} = t\,(r + \beta PV(\Gamma))_{\hat{\rho}(k,t)}.$$

Consequently,

$$\max Y(k, t) \geqslant V_k(\Gamma) + t \max (r + \beta PV(\Gamma))_{\hat{\rho}(k,t)}$$

and the second term in the right hand side of this inequality is at least $tV_k(\Gamma)$, by lemma 3.1. Hence

$W_k(t+1) = \min \{\max (t+1)^{-1}\ Y(k, t), W_k(t)\} \geqslant \min \{V_k(\Gamma), W_k(t)\} = V_k(\Gamma)$,

for each $k \in S$. This completes the proof of the lemma. □

Proof of theorem 3.1: Take $k \in S$. Denote $\lim_{t\to\infty} W(t)$, which limit exists, in view of lemma 3.1, by W^*. For $t \in \mathbf{N}$, let $A_k(t)$ be the $n_1(k) \times n_2(k)$-matrix, with in the (i, j)-th cell the real number

$$a_{ij}(t) = r(k, i, j) + \beta \sum_{\ell=1}^{N} p(\ell \mid k, i, j)\, W_\ell(t).$$

Now $\lim_{t\to\infty} A_k(t)$ exists by lemma 3.1, and equals

$$A_k = [r(k, .\,, .) + \beta \sum_{\ell=1}^{N} p(\ell \mid k, .\,, .)\, W^*_\ell].$$

It is obvious that

$\langle X(k, t), Y(k, t)\colon t \in \mathbf{N}_0 \rangle$ is a vector system for the sequence

$\langle A_k(t)\colon t \in \mathbf{N} \rangle$.

Hence, by theorem 2.1, we have

$$\lim_{t\to\infty} t^{-1} \max Y(k, t) = \lim_{t\to\infty} t^{-1} \min X(k, t) = \mathrm{val}(A_k).$$

Taking the limit in (S.1), and using the above equality, we obtain

$$W^*_k = \min \{\mathrm{val}(A_k), W^*_k\} \text{ or}$$

$$W^*_k \leqslant \mathrm{val}\,[r(k, .\,, .) + \beta \sum_{\ell=1}^{N} p(\ell \mid k, .\,, .)\, W^*_\ell].$$

This inequality holds for each $k \in S$. From lemma 1 in *Denardo/Fox* [1968, p. 474], it follows then that $W^* \leqslant V(\Gamma)$. Conversely, from (3.1) in the proof of lemma 3.2, we obtain $W^* \geqslant V(\Gamma)$. Now $W^* = V(\Gamma)$ implies, by lemma 3.1, that val $(A_k) = V_k(\Gamma)$. Hence, theorem 3.1 is proved. □

Proof of theorem 3.2: By theorem 2.2, $\hat{\pi}(k)$ and $\hat{\rho}(k)$ are optimal mixed actions in the matrix game $A_k = [r(k, .\,, .) + \beta \sum_{\varrho=1}^{N} p(\varrho \mid k, .\,, .)\, V_\varrho(\Gamma)]$ for player 1 and player 2, respectively. This holds for each $k \in S$. The theorem follows now from lemma 3.1. □

4. Remarks

(4.1) The vector $W(t)$, introduced in section 3, approximate $V(\Gamma)$ from above. If one wants also an approximation from below, then one can modify the iteration procedure, by starting with vectors $X(k, 0)$, $Y(k, 0)$ and $W'_k(0)$ with

$$W'_k(0) = \min X(k, 0) = \max Y(k, 0) \leqslant V_k(\Gamma),$$

and where (S.1) is replaced by

$$(\text{S.1})'\ W'_k(t) = \max \{\min t^{-1}\ X(k, t-1),\ W'_k(t-1)\},$$

and $W(t)$ in (S.2) and (S.3) is replaced by $W'(t)$. Then $[W'_k(t), W_k(t)]$ is an estimation interval around $V_k(\Gamma)$, which length shrinks to zero, when t increases to infinity.

(4.2) Let $B(S, \{A_{mk}\})$ be the family of all stochastic games with S as state space and the A_{mk} as action spaces.

Let $U(S, \{A_{mk}\})$ be the subset of $B(S, \{A_{mk}\})$, consisting of those games with unique optimal stationary strategies.

In *Tijs/Vrieze* [1980, theorem 3.4] asserts that $U(S, \{A_{mk}\})$ is an open and dense subset of $B(S, \{A_{mk}\})$. Suppose now, that $\Gamma \in U(S, \{A_{mk}\})$. Then we can sharpen theorem 3.2 as follows: the sequences

$$\langle \hat{\pi}(1, t), \hat{\pi}(2, t), \ldots, \hat{\pi}(N, t) \rangle \text{ and } \langle \hat{\rho}(1, t), \hat{\rho}(2, t), \ldots, \hat{\rho}(N, t) \rangle$$

converge to the unique optimal stationary strategies of player 1 and player 2, respectively.

(4.3) Many modifications of the Brown-Robinson method were considered, such as alternating moves instead of simultaneous moves, weighting of past choices etc. Of course, such modifications can also be made without many difficulties in the case of converging sequences of matrix games and for stochastic games.

(4.4) Let us look at a stochastic game Γ, where in each stage one player, say player

2, is a dummy i.e. for each $k \in S$, $A_{2k} = \{1\}$. Then the remaining problem equals a Markov decision problem. Suppose, w.l.o.g., that all payoffs are negative (which can be realized by substracting a large constant c from each payoff, resulting in a decrease of the total discounted payoff by $(1-\beta)^{-1}c$). Take $Y(k, 0) = 0 \in \mathbf{R}^{n_1(k)}$, $X(k, 0) = 0 \in \mathbf{R}$. Then the algorithm yields:

$$t^{-1} y_i(k, t) = r(k, i, 1) + \beta \sum_{\ell=1}^{n} p(\ell \mid k, i, 1) \tilde{W}_\ell(t),$$

where $\tilde{W}_\ell(t) = t^{-1} \sum_{i=1}^{t} W_\ell(\tau)$. This scheme shows some similarity with the successive approximation methods for solving Markov decision problems [cf. *Howard*], but the convergence rate may be slower. In Markov decision theory, use is made of the iteration scheme

$$W_k(t) = t^{-1} \max Y(k, t)$$

and

$$t^{-1} y_i(k, t) = r(k, i) + \beta \sum_{\ell=1}^{n} p(\ell \mid k, i) W_\ell(t-1),$$

which assures a geometric rate of convergence. However, in a similar way as for this algorithm, it can be shown for our algorithm, that for t sufficiently large, only actions $i(k, t)$ are chosen, for which $\pi_1(t) = (i(1, t), \ldots, i(N, t))$ is an optimal stationary strategy. Moreover, if we change our iteration scheme a bit, by taking

$$W_k(t) = \min \{\max t^{-1} Y(k, t-1), W_k(t-1), \max (r + \beta P W(t-1))_{j(k,t)}\}$$

then, in the case of Markov decision problems, our algorithm and the above mentioned algorithm for Markov decision problems give the same iteration scheme. For stochastic games, also in the case where player 1 has an optimal pure stationary strategy, this modification of the iteration scheme gives an improvement of the convergence rate, which becomes geometric then. But for general stochastic games no improvement might be expected.

(4.5) *Shapiro* [1958] obtained an a priori estimate of the convergence rate of $O(t^{-(1/m+n-2)})$ of the Brown-Robinson scheme for solving an $m \times n$-matrix game. In a similar way as in that paper, we can prove that for a stochastic game Γ, for each $k \in S$:

$$\max t^{-1} Y(k, t) - \min t^{-1} X(k, t) \leqslant K 2^{n_1(k)+n_2(k)} t^{-(1/n_1(k)+n_2(k)-2)},$$

where $K = (1-\beta)^{-1} \max \{|r(k, i, j)| : (k, i, j) \in T\}$.

Hence, in spite of the fact that, during the iteration, also the limit matrix is approximated, the a priori convergence rate is not worse than in the Brown-Robinson procedure. This shows that the possibilities of our iteration method of being a competitor of the usual successive approximation methods is much greater than the possibilities of the Brown-Robinson procedure for matrix games in comparison to e.g. linear programming. This may be illustrated by figure 1.

	LP/Succ. approx.	Brown-Robinson/our algorithme
Matrix game of size $m \times n$	to solve 1 dual pair of LPP of size $m \times n$	At each step to compute the maximal coordinate of an m-vector and the minimal coordinate of an n-vector.
Stochastic game of size $n_1(k) \times n_2(k)$ in state k	at each step to solve N dual pairs of LPP of size $n_1(k) \times n_2(k)$; $k = 1, \ldots, N$.	at each step to compute the maximum of an n_1 (k)-vector and the minimum of an n_2 (k)-vector for $k = 1, \ldots, N$.

Fig. 1

References

Brown, G.W.: Some Notes on Computation of Game Solutions. RAND Report P-78, The RAND Corporation, Santa Monica, California 1949.

–: Iterative Solution of Games by Fictitious Play. Activity Analysis of Production and Allocation. New York 1951, 374–376.

Danskin, J.M.: Fictitious play for continuous games. Naval Research Logistics Quarterly **1**, 1954, 313–320.

Denardo, E., and *B. Fox*: Multichain Markov renewal programs. SIAM J. Appl. Math. **16**, 1968, 468–489.

Dresher, M.: Games of Strategy; Theory and Applications. Englewood Cliffs 1961.

Gale, D.: The Theory of Linear Economic Models. New York 1960.

Howard, R.: Dynamic Programming and Markov processes. New York 1960.

Karlin, S.: Mathematical Methods and Theory in Games, Programming and Economics. Reading 1959.

Luce, R.D., and *H. Raiffa*: Games and Decisions. Introduction and critical survey. New York 1957.

McKinsey, J.C.C.: Introduction to the Theory of Games. New York 1952.

Miyasawa, K.: On the Convergence of the Learning Process in a 2 × 2 Non-Zero-Sum Two-Person Game. Research Memorandum **33**, Princeton 1961.

Owen, G.: Game Theory. Philadelphia 1968.

Rauhut, B., N. Schmitz and *E.-W. Zachow*: Spieltheorie. Stuttgart 1979.

Robinson, J.: An iterative Method of Solving a Game. Ann. of Math. **54**, 1950, 296–301.

Rosenmüller, J.: Über Periodizitätseigenschaften spieltheoretischer Lernprozesse. Z. Wahrscheinlichkeitstheorie verw. Geb. **17**, 1971, 259–308.

Shapiro, H.N.: Note on a Computation Method in the Theory of Games. Comm. on Pure and Appl. Math. **11**, 1958, 587–593.

Shapley, L.S.: Stochastic Games. Proc. Nat. Acad. Sci. USA **39**, 1953, 1095–1100.

–: Some Topics in Two-Person Games. Annals of Math. Studies **52**, 1–28, Princeton 1964.

Tijs, S.H., and *O.J. Vrieze*: Perturbation Theory for Games in Normal Form and Stochastic Games. J. Opt. Th. and Appl., 1980, 549–567.

Van den Akker, A.G.: Some subjects in game theory; Markov games, continuous games (in Dutch). Master's thesis, University of Technology, Eindhoven, The Netherlands, 1976.

Van der Wal, J.: Discounted Markov Games; Successive Approximation and Stopping Times. Int. Journal of Game Theory **6**, 1977, 11–22.

Received January 1980